WPS4149

The Worldwide Governance Indicators Project: Answering the Critics

Daniel Kaufmann, Aart Kraay, and Massimo Mastruzzi[1]
The World Bank

Abstract: The Worldwide Governance Indicators, reporting estimates of six dimensions of governance for over 200 countries between 1996 and 2005, have become widely used among policymakers and academics. They have also attracted some explicit written criticisms. In this short paper we synthesize eleven critiques offered by four recent papers. We then refute them as either conceptually incorrect or empirically unsubstantiated.

World Bank Policy Research Working Paper 4149, March 2007

[1] 1818 H Street NW, Washington, DC 20433, dkaufmann@worldbank.org, akraay@worldbank.org, mmastruzzi@worldbank.org.

In this paper we summarize and respond to some recent critiques of our Worldwide Governance Indicators project (as described in our series of papers Kaufmann, Kraay and Zoido-Lobatón (1999a,b) and (2001), and Kaufmann, Kraay and Mastruzzi (2004, 2005, and 2006)). The latest round of these governance indicators reports on six dimensions of governance every two years since 1996, and annually between 2002 and 2005, for over 200 countries. For brevity we will refer to them here as the WGI. The WGI are based on the aggregation of perceptions of governance from 31 different data sources provided by 25 different organizations, and so provide a synthesis of the views of a very large and diverse group of stakeholders regarding the quality of governance across countries. We report the aggregate governance indicators, the underlying individual indicators from all but three of the sources, together with accompanying descriptive papers and a Web-based interactive data tool, at www.govindicators.org.

While this paper is primarily devoted to responding to critiques of the WGI, we think it useful to begin by noting that the WGI have in recent years become among the most widely-used indicators of governance by policymakers and academics.[2] The usefulness of the aggregate indicators in the WGI stems from the fact that (a) they provide very broad country coverage, greater than that provided by any individual data source on governance; (b) by averaging information from many different data sources they are able to conveniently summarize the wealth of existing information on governance; (c) by averaging they are also able to smooth out some of the inevitable idiosyncracies of individual measures of governance and so be more informative about the broad notions of governance they are intended to measure than any individual data source; and (d) the estimates of governance are (unusually in this field) accompanied by explicit margins of error that transparently indicate the unavoidable degree of uncertainty associated with measuring governance by any means. Indeed, we think that it is in part because of the widespread use of the WGI that they are increasingly also beginning to attract critiques. We think that this process of discussion and debate of these critiques is very useful in identifying -- but also often discarding -- potential problems that arise in efforts to measure governance.

[2] For example, the United States Millennium Challenge Account aid program prominently relies on five of the WGI in its procedures for determining country eligibility, see www.mcc.gov for details.

Here we address eleven specific criticisms of the WGI that are made in four recent papers (Arndt and Oman 2006 (AO), Knack 2006 (K), Kurtz and Shrank 2006 (KS), and Thomas 2006 (T)). AO provide an extensive and very useful survey of the many different types of governance data available, and in fairness we note that AO have many nice things to say about the WGI, kindly referring to them as "probably the most carefully constructed governance indicators". Here we focus only on addressing their main criticisms, contained in Section 4 of their paper. Similarly, K's focus is on interpreting the available data on trends in corruption in countries in Europe and Central Asia (ECA) between 2002 and 2005, and also contains in our view a very useful and thorough review of the many different types of data available to measure corruption in these countries. However, along the way he raises several criticisms of the WGI as well as the underlying data with which we disagree. The paper by KS, which is forthcoming in the Journal of Politics, is primarily focused on critiquing the WGI. We have prepared a fuller response to the issues they raise for publication in the same journal, and we refer the interested reader to that article for details. Finally, T's paper is also a critique specifically of our indicators, and we respond to it here.

We organize the points made in these papers into eleven related critiques, and provide our responses. The first four critiques call into question the usefulness of the Worldwide Governance Indicators for making comparisons of governance over time and across countries. Critiques 5 and 6 allege various sorts of biases in the individual indicators underlying our aggregate governance indicators. Critiques 7 and 8 concern the independence of the assessments of governance provided by our different data sources, and the consequences for the aggregate governance indicators. Critique 9 responds specifically to the main thesis of T that the WGI are an "elaborate untested hypothesis" because we fail to provide evidence of "construct validity", a somewhat obscure term that we define below. Critique 10 deals with concerns primarily of T regarding access to the data used in the WGI. Finally, the 11th critique raised by AO, while not specifically about the WGI, refers to a paper of ours on the causality between governance and growth that used data from the WGI.[3]

[3] KS also criticize the literature on governance and growth more broadly, and we provide a response, in our forthcoming Journal of Politics article,

In short, we do not find the criticisms raised in these four papers to be particularly compelling. As we argue below they are usually based on misinterpretations of our indicators, or of the empirical evidence involving these indicators.[4] Moreover, we note that many of the concerns raised by our critics are fairly generic and so would apply to many other types of individual and aggregate governance indicators, and not just the WGI project. We highlight such cases below.

<u>Critique 1:</u> Governance cannot be compared over time using the WGI since they are scaled to have the same global averages in every period

Variants on this critique are raised by both AO and K. AO first correctly point out that our aggregate governance indicators are scaled to have a zero mean and unit standard deviation in each period. They then go on to assert that this means that the WGI "...cannot reliably be used for monitoring changes in levels of governance over time, whether globally, in individual countries, or among specific groups of countries" (AO 2006, p. 61). With the exception of global averages, this statement simply is incorrect. We have clearly acknowledged in our past work that by setting the world average of governance to zero in each period, our aggregate indicators are obviously not informative about trends in global averages of governance by definition. Recognizing this, we have in the last three updates of our indicators also provided whatever evidence we could from a selection of our individual underlying sources that are consistently available for longer periods of time about trends in world averages (see for example Kaufmann, Kraay and Mastruzzi 2004, Table 7; 2005, Table 6; and 2006, Table 5). These exercises have turned up little evidence of significant trends in world averages of governance, and so our choice of units for governance which sets the world average to zero in each period is innocuous.

This evidence from our individual sources that world averages of governance are not changing much is crucial, because it allows us to interpret the *relative* changes in country scores on our aggregate indicators, or groups of countries' scores, as *absolute* changes. In particular, if world averages do not change, then it is appropriate for us to

[4] We have privately responded to the authors of each of the papers, presenting our views on these issues along the lines presented below. The purpose of this note is to place our responses to these criticisms in the public domain

rescale our governance indicators to have the same mean in each period, and there is no difference between changes in countries' relative positions on our indicator, and their absolute changes. This point has been made in Kaufmann, Kraay, and Mastruzzi (2004, 2005, and 2006).

K raises a more sophisticated objection to our normalization of a zero mean and unit standard deviation. Referring to our corruption indicator, he notes that the country coverage of this indicator has increased substantially over time (in fact, from 152 countries in 1996 to 204 countries in 2005). He then correctly notes that adding new countries can in principle change the ranks of existing countries in a relative ranking like our corruption indicator. If for example we add a country with very low corruption, i.e. a very high score on our Control of Corruption indicator, then this will reduce the rank of all of the other countries in the sample by construction. While technically this point is correct, practically it turns out to not to matter much, for three reasons:

- As a minor point, K's primary interest is in trends in corruption in the ECA region between 2002 and 2005. During this period the country sample covered by our Control of Corruption indicator changes inconsequentially, from 197 to 204 countries worldwide. Even if all seven countries added had the lowest corruption in the world (and they do not), this would lower the percentile rank of the remaining countries by only about 3 percent, which is well within the margins of error for changes in country scores that we continually encourage users to take into account when making comparisons of changes over time.

- If users are interested in comparing the relative ranks of a set of countries over time (for example, the ECA countries relative to the rest of the world as is the case in K), then this problem can trivially be circumvented by simply looking at country ranks in a common set of comparator countries in both periods. This requires nothing in the way of sophisticated statistical tools, nor does it require access to the underlying data. This point is also noted by K.

- The extent to which the addition of new countries affects the relative ranks of countries already in the sample depends on how different the "entrants" to the sample are relative to the "incumbents". In the case of corruption between 1996

and 2005, the difference appears not to be very large. One way to check this is to look at the mean value of our 2005 Control of Corruption measure for the "entrants" between 1996 and 2005. The mean score of the entrants is 0.06 (or just about 1 percent of the range of the indicator), which is only slightly, and not significantly, above the world average which by construction is set equal to zero. In simple terms, this means that the new countries added to the sample included some with very little corruption, and some with a lot of corruption, and so the ranks and scores of the remaining countries are not systematically affected by the addition of these countries.

In fairness, however, it does turn out that for some of our other indicators, there are a bit bigger -- but never large -- differences in the mean scores of the "incumbents" and "entrants" over different subperiods. In the 2005 update of the governance indicators, we have provided small adjustments to the mean and standard deviation of the aggregate governance indicators for earlier periods to recognize this changing sample composition. In short, we maintain our assumption (as a defensible choice of units) that governance has a zero mean and unit standard deviation across all countries in the world in each period. However, recognizing that in earlier years we do not have data for all countries, we allow the mean and standard deviation of the governance indicators to be slightly different from zero and one respectively *in the sample of countries for which we actually have data*. For details refer to Section 2 of Kaufmann, Kraay and Mastruzzi (2006).

<u>Critique 2:</u> Governance cannot be compared across countries or over time with the WGI since the estimates for governance for different countries or periods may be based on different underlying data sources.

K and AO also both raise, in varying detail, this issue that arises in making comparisons over time and across countries. They both correctly note that when comparing two countries (or periods) using one of our governance indicators, the estimates of governance for the two countries (or periods) might be based on two different, and, in a few extreme cases, even on wholly non-overlapping, sets of underlying data sources. While this point is factually correct of course, we do not find it to be a serious criticism, for five reasons:

- In the extreme case where two countries of interest do not appear in any single common data source, we actually would argue that one of the strengths of our aggregate indicators is that they do in fact make it possible to compare governance in these countries, despite the lack of common sources. After all, one way to think about our aggregation methodology is that it provides a reasonably sophisticated way of placing very different underlying data sources into common units, and this is precisely what permits comparisons across countries that do not appear in the same sources. To take a specific example, suppose one wanted to compare corruption in the Bahamas with Saint Kitts and Nevis in 2000. This is a quite unusual case where there are two countries appearing in non-overlapping sets of underlying sources.[5] In particular, these two very small countries each happen to appear in only one of our data sources for corruption in that earlier year, the Bahamas in the ICRG ratings produced by Political Risk Services, and Saint Kitts and Nevis in the World Bank's CPIA ratings. The virtue of our aggregate indicator is that it provides a way of putting the scores from these very different agencies into common units and permits comparison between them despite the absence of a common data source, subject of course to the margins of error that we report, and that would be large for such countries that unusually appear in only one data source each. Admittedly this is an extreme example, but a more general point holds: if we want to make comparisons between countries based on a common set of data sources, this limits the information set we have available on which to base our judgments (in the extreme case of these two Caribbean states, it would eliminate the information set completely and prevent any comparison). In fact, one of the motivations we originally had in constructing the WGI was to enable comparisons across as large a set of countries as possible.

[5] While useful for illustrative purposes, this example is highly unusual in two respects. First, the two countries appear in only one data sources, while in 2005, only 15 of the 204 countries covered by the Control of Corruption indicator appear in only data source. Second, the fact that these two countries share no common data sources is even more unusual. Looking across our six aggregate indicators, only about one percent of all possible pairwise country comparisons involve countries with no common data sources. By contrast, roughly half of all possible pairwise comparisons are based on *at least* five common data sources.

- Related to the previous point about units, we disagree with K's argument that since each underlying data source measures a somewhat different concept of corruption, the implicit definition of corruption is different when we compare two countries with different sets of underlying data sources. Again, it is useful to think about our aggregation method as a way of putting different data sources in common units. Suppose one data source measures corruption in procurement, while another measures corruption in the judiciary, and suppose once again we want to compare one country that appears only in the one indicator with another that appears only in the other indicator. Our aggregate indicator extracts the common component of these (and all our other data sources), which we label as overall "Control of Corruption". That is, we have just one implicit definition of corruption, which comes from the aggregation of these many data sources across many countries. Using the aggregate indicator we of course cannot distinguish between these particular dimensions of corruption, and for policy purposes in a particular country this distinction may be useful.[6] But what our indicator does do is allow us to compare the extent of overall corruption in the two countries, based on the imperfect information both particular indicators provide about overall corruption.

- Whether this criticism is practically important or not depends a lot on whether (a) different data sources successfully distinguish between different dimensions of corruption, and (b) the different nuances of corruption measured by different sources really differ a lot across countries. If for example some countries have very "clean" judiciaries but high administrative corruption, while other countries are the other way around, and if data sources were able to sharply distinguish between the two, then a measure of overall corruption based on measures of administrative and judicial corruption would not be very informative. We note first that several of the individual data sources in fact have quite general questions about corruption. And in cases where a single data source distinguishes between alternative forms of corruption, we in fact average together the different

[6] This is something we have long acknowledged. For example, in our very first paper, Kaufmann, Kraay and Zoido-Lobatón (1999a) we write in the conclusion that "There is therefore a need to improve the quality and quantity of governance data, both by improving and extending cross-country survey work of governance perceptions, as well as employing country-specific in-depth governance diagnostics", and similar statements can be found in our subsequent updates of the governance indicators.

questions before including them in our aggregate indicators. So we are not of the view that the definitional distinctions across our data sources are in fact very large. Moreover, it turns out that even questions about ostensibly different forms of corruption tend to be very correlated with each other. In simple terms, it seems unlikely that there would be many countries with high judicial corruption but low administrative corruption, and vice versa. And this is what the data tells us. For example, for Control of Corruption in 2005, the median correlation of our 18 underlying data sources with the aggregate indicator is 0.85, and only two data sources are correlated at less than 0.5.[7]

- Related to the previous point, in past work we have empirically documented the extent to which changes over time in our aggregate governance indicators for individual countries are influenced by the addition of data sources. In principle, of course, the addition of a new data source for a country that provides a very different assessment than other data sources can result in a large change in the aggregate indicator for that country. In practice, however, this effect does not appear to be very important, and accounts for just a small portion of the variation over time in country scores. In Kaufmann, Kraay, and Mastruzzi (2005) for example, we looked at all "large" -- in the sense of being statistically significant -- changes in each of our six governance indicators between 1996 and 2004. We first computed what the change in our estimate of governance would have been based on a common set of indicators, and then isolated the remaining component of the change which reflected the addition of data sources for each country. On average, we found that the addition of data sources accounted for only about 9 percent of the variation in changes in our aggregate indicators, for countries with large changes in governance.

[7] These two are the Latinobarometro survey of countries in Latin America, and Freedom House's Countries and the Crossroads report. One other source with a low correlation for corruption is the BEEPS survey (correlation with aggregate indicator of only 0.55. Since one of the main interests of K is to account for differences between the BEEPS survey and other measures of trends in corruption in the ECA region, we are sympathetic with K's emphasis on the differences among data sources for this particular indicator and region. However, we do not think the point is more generally true for the majority of our data sources, which tend to be very much in agreement with each other.

- Fifth and finally, a simple practical point. For many purposes we do recognize that it can be of interest for some users to make comparisons of governance based on particular individual data sources. To facilitate this, with the 2005 release of the aggregate governance indicators, we have made all but three of our underlying data sources available to users on our website.[8]

<u>Critique 3:</u> Changes over time in some of the individual indicators underlying the WGI aggregate indicators reflect corrections of past errors rather than actual changes.

This critique of several of our underlying indicators is made by K in the context of his discussion of trends in governance in countries in the ECA region between 2002 and 2005 . We do not think that the evidence provided by K supports his claim. K argues that in many cases individual data sources change their assessments of governance in a country simply to correct past errors in their assessments: countries that were rated "too high" in the past get lower scores, and vice versa. K then goes on to document that among ECA countries there is "regression to the mean", in the sense that changes in governance tend to be negatively correlated with initial scores, and interpret this as evidence that risk rating agencies change their scores to correct past "errors".

We believe that this is an overinterpretation of his results based on a regression that has a much simpler explanation. To be specific, let y(j,t) be the rating of country j in year t. K regresses y(j,t)-y(j,t-1) on y(j,t-1), which is of course mathematically identical to regressing y(j,t) on y(j,t-1): the coefficient on the initial value in the first regression will just be the coefficient in the second regression, minus one. But the coefficient in the second regression, which is just the autocorrelation coefficient of the governance rating, is for most of these data sources a positive number between 0 and 1. Suppose next that the governance rating is a noisy proxy for true governance, i.e. that y(j,t) = g(j,t) + e(j,t) where g(j,t) is true governance and e(j,t) is the error made by a particular source. It seems quite plausible to us that governance on average changes rather slowly over

[8] These three sources are the Country Policy and Institutional Assessments produced by the World Bank, the African Development Bank, and the Asian Development Bank, which for the most part are treated as confidential by these organizations. Only in the past few years has limited, but growing disclosure of this data been made by these organizations, and full public access to the detailed disaggregated and historical data on which we rely is still not permitted..

time, indicating that g(j,t) would have strong persistence, which would be reflected in strong persistence in y(j,t). Thus, as long as governance is persistent, e.g. it tends to generally change only slowly over time, we should expect to find a negative coefficient in K's regression reflecting nothing more than such persistence in governance itself**.**

We thus argue that K provides no direct evidence in support of his claim that changes in governance estimates reflect "correction" of past mistakes. We do think however that it may be plausible *a priori* that similar kinds of corrective mechanisms could be at work in some of our indicators. For example, in our latest update of the governance indicators we have devised a test of the hypothesis that data sources update their scores in order to reduce past discrepancies between themselves and other data sources (Kaufmann, Kraay, and Mastruzzi (2006), Section 3). The simple intuition is that if data sources update their scores to reduce the past differences between them and other data sources, we should expect to see that the different data sources become more correlated with each other over time. K provides some evidence that this is the case, but only for measures of corruption in ECA countries over the past few years which is his primary interest. But it would be wrong to conclude from this that it is a general pattern. We have examined trends over time in the pairwise correlations between our sources for all countries, between 1996 and 2005, and find no systematic evidence of increased correlations (Kaufmann, Kraay and Mastruzzi 2006, Table 7, and accompanying discussion). The median change in correlation is only 0.03, and roughly the same number of pairs of sources exhibit increased and decreased correlations over time. We therefore do not think that this kind of updating is empirically very important.

<u>Critique 4:</u> The WGI are too imprecise to permit meaningful comparisons of governance over time or across countries

AO argue at some length that the WGI "...do not allow for a reliable comparison of levels of governance over time..." (AO 2006, p. 68). The gist of their critique is that, since only a relatively small number of countries experience changes in governance that are large enough to be considered statistically significant, the indicators cannot be used to make comparisons over time. We find this critique peculiar, and entirely without basis. First, it is not clear to us how the fact that many countries do not experience

significant changes in governance according to our measures is a drawback of the WGI.[9] Absent other information that governance in such countries is indeed changing but our indicators miss the changes, or conversely, without evidence that governance is indeed not changing in countries where our indicators show significant changes, AO's assertion is purely speculative. The presence of margins of error in our indicators does not obviate the ability to make comparisons over time -- rather it enhances the ability of the user to make comparisons over time, by providing guidance as to which observed changes are likely to be meaningful, and which are not. In fact, we would argue that the presence of explicit margins of error in the WGI serves as a useful antidote to the type of superficial "elevator" discussions of governance that are unfortunately common -- this country went up, that one went down, a third is the best in the world, another is the worst in the world etc. Such discussions, without due regard to the limitations of the data as expressed in the margins of error, which apply to any data source on governance or investment climate, are not very informative.

With respect to cross-country comparisons, we have always encouraged users of the governance indicators to take margins of error into account when making comparisons across countries. But this encouragement does not mean that no significant comparisons can be made. Consider for example our Control of Corruption indicator in 2005 which covers 204 countries, so that it is possible to make 20,706 pairwise comparisons of corruption across countries using this measure. For 64 percent of these comparisons, 90% confidence intervals do not overlap, signaling quite highly statistically significant differences across countries. And if we lower our significance level to 75 percent, which may be quite adequate for many applications, we find that 74 percent of all pairwise comparisons are statistically significant. While we continue to emphasize to users that many of the small differences between countries may well be neither statistically or practically significant, we also emphasize that a great many significant differences between countries can in fact be established using our aggregate indicators. Indeed, we reiterate that only by using aggregate indicators with transparently-reported margins of error (such as the WGI) are users even able to know whether observed differences in point estimates of governance are in fact significantly

[9] And indeed, such a criticism would in principle also apply to any other measure of governance, were it not for the fact that these other measures do not explicitly acknowledge their margins of error and so fail to distinguish between significant and insignificant changes.

different across countries. The vast majority of existing governance data sources do not report such margins of error, even though measurement error is surely present in them as well, and so it is difficult for users to assess the significance of differences across countries or over time.

<u>Critique 5:</u> The individual indicators underlying the WGI are biased towards the views of business elites, and thus so are the aggregate indicators.

This concern is raised by both AO and KS. It is apparently based on the observation that several of our data sources are commercial risk rating agencies (whose main clients are businesses), as well as a number of cross-country surveys of firms. In short, they argue, businesspeople like low taxes and minimal regulation, while the public interest demands reasonable taxation and appropriate regulation. Estimates of governance based on the perceptions of businesspeople, and especially the "elite" among businesspeople, will therefore necessarily be biased. We do not think this criticism is particularly valid, for three broad reasons.

- First, we note that our indicators rely on much more than just the views of businesspeople. In the latest 2005 update of our governance indicators, our data sources include four cross-country surveys of firms, as well as seven commercial risk rating agencies, which one might expect to reflect narrower business interests. But we also rely on three cross-country surveys of individuals, six sets of ratings produced by government and multilateral organizations (such as the World Bank, the African Development Bank and the US State Department), and finally another 11 data sources produced by a wide range of non-governmental organizations (such as Freedom House, Reporters Without Borders, and many others). It is therefore simply incorrect to dismiss our indicators as reflecting solely -- or even primarily -- the narrow interests of the business elite.

- Second, it is not at all the case that firm surveys focus exclusively on either foreign investors in a country, or else the "elite" of large domestic firms. In fact, in the largest cross-country survey of firms that we use, the Global Competitiveness Report, just 14 percent of respondent firms are foreign-owned in the 2005 round of the survey. Moreover, 30 percent of all respondent firms are

quite small with less than 50 employees, and 43 percent of firms have less than 100 employees. In contrast only 19 percent of firms are very large with employment greater than 1000. Are these figures truly representative of the size distribution of firms? This is hard to know because information on the size distribution of all firms is very hard to obtain in a systematic way across countries. We do have some limited evidence on the distribution of all registered firms for EU countries for 2000.[10] In a sample of 32 EU countries, 24 percent of firms fall in the 10-50 employee size category. In the GCS for these countries, it turns out that exactly the same fraction of firms have employment less than 50. This exaggerates the representativeness of the GCS though because the EU data also report employment for very small firms with less than 10 employees, which account on average for a very long tail some 40 percent of firms. But setting aside these very small firms, the GCS does not appear to be broadly skewed towards large or otherwise "elite" firms.

- Third, the extent to which this critique is valid depends crucially on the extent to which there are fundamental differences between the perceptions of business people and those of other members of society as to what constitutes good governance. If this is true, then the responses of firms (or commercial risk rating agencies who serve mostly business clients) to questions about governance should not be very correlated with ratings provided by respondents who are more likely to sympathize with the common good, such as individuals, NGOs, or public sector organizations. In fact, overall there are quite strong correlations among most of our difference types of data sources. As an example, consider the ingredients of our Government Effectiveness indicator for 2005. The correlation between two of our major cross-country firm surveys is 0.74, and the correlation of these firms surveys with a survey of households in Africa is very similar at 0.7. More systematically, as we discuss further below, the rankings provided by our aggregate indicators are quite robust to alternative weighting schemes. This robustness reflects precisely the fact that on average our different types of data sources provide highly correlated assessments. This in turn suggests to us that it is implausible that the preferences of businesspeople regarding good governance differ so dramatically from those of other types of respondents.

[10] We are grateful to Leora Klapper for providing this data.

A related concern, not explicitly raised by the critics we address here, but commonly heard nevertheless, is that expert assessments are not just biased (possibly to the interests of business elites), but rather that the experts simply get things wrong. In a recent paper Razafindrakoto and Roubaud (2006) use specially-designed surveys in eight African countries to contrast corruption perceptions based on household surveys with those based on expert assessments. The unique feature of this exercise is that the experts were asked to predict the country-level average responses from the household survey. In this sample of eight countries it turns out that the experts' ratings were essentially uncorrelated with the household survey responses. Razafindrakoto and Roubaud (2006) conclude that the household surveys capture the "objective reality" of petty corruption and that the experts are just plain wrong. While this is a very creative and interesting effort, we disagree with their conclusion for two reasons.

- First, it is not at all clear why there should be measurement error only in the expert assessment and not in the household survey. Households were asked whether they had been a "victim of corruption". There are a variety of reasons why households might think they were victimized by corruption when in fact it was not present. For example, a patient waiting in the queue to see a state-provided doctor might think (incorrectly) that people at the head of the queue had bribed someone to get there. Conversely households might well have paid a bribe, received the associated benefit, and found themselves quite satisfied and not at all "victimized" by the transaction. Our rather more modest interpretation of their finding is that there is measurement error in any estimate of corruption, regardless of the identity of the respondent.
- Second we note that the low correlation between expert and household survey responses does not seem to be more broadly true in larger samples of countries. As pointed out elsewhere in this note we have found many cases where household of firm survey responses about corruption and other dimensions of governance are highly correlated with those of expert assessors. And we note also that correlations among different household survey measures of corruption are not particularly high, as would be the case if household surveys are the most reliable way to measure corruption.

<u>Critique 6</u>: *The data sources underlying the WGI are overly influenced by recent economic performance, and/or the level of development of a country -- rich, or fast-growing countries get better scores simply because they rich or growing fast.*

This critique is a common one, and is made at length by KS, and also in passing in another widely cited paper, Glaeser et. al. (2004). The gist of the argument is simple. Governance ratings, especially those produced by commercial risk rating agencies, assume that governance must be good in countries that are rich or enjoying recent strong economic performance, and so these countries receive ratings that are better than they deserve. This phenomenon is sometimes referred to as "halo effects", and is something that we have studied in our earlier work with these indicators.

- In Kaufmann, Kraay, and Mastruzzi (2004) we look for evidence of halo effects associated with levels of development. Glaeser et. al. (2004) argue that much of the observed correlation between governance and levels of development can be explained by such halo effects. We develop a simple statistical model to investigate the empirical importance of this claim (which Glaeser et. al. do not), and show that there is in fact a tradeoff Halo effects can be thought of as measurement error. By itself, greater measurement error in governance actually *lowers* the correlation between governance and per capita incomes, while measurement error that is correlated with per capita incomes *increases* it. Given this tradeoff, we provide calibrations that show that halo effects would have to be implausibly strong in order to account for the observed high correlation between governance and per capita incomes.

- In contrast, KS do claim to provide direct empirical evidence of halo effects, showing that one of our six governance indicators, Government Effectiveness, tends to have a significant partial correlation with two-year average growth rates prior to the date of the governance indicator in a limited set of regressions that they report. In our dedicated full response to KS, forthcoming in the Journal of Politics, we document in detail how the evidence they report is not robust, and in any case is misinterpreted by KS. In brief, we show that very minor changes to their empirical specification entirely overturn their results. We also show that after controlling for long-run economic performance of countries, the short-term

growth that KS claim is driving halo effects is also no longer significant. Based on this we argue that the short-run growth variable is simply proxying for longer-run growth, and that the KS regressions could just as well be interpreted as picking up an entirely reasonable causal effect of good governance on long-run economic performance. Consistent with this, we show that a very careful measure of government effectiveness that KS -- likely correctly -- hold up as a model indicator untainted by "halo effects" exhibits the same partial correlations with long- and short-run growth as do the WGI. We therefore do not find their evidence of alleged "halo effects" to be at all compelling.

<u>Critique 7:</u> The individual data sources underlying the WGI, particularly those from commercial risk rating agencies, make correlated errors in their assessments of governance, and thus are less informative about governance than they appear.

This criticism (together with Critique 8 below) is discussed at length in AO (pp. 65-67), as well as in K (pp 21-27)). The point here is a simple one. Suppose that one cross-country rating agency "does its homework" and comes up with an assessment of governance for a set of countries based on its own independent research, but a second rating agency simply reproduces the assessments of the first. Then in reality we would only have one data source, not two, and inferences about governance based on the two data sources would be no more informative than inferences based on just one of them. In short, the rationale for constructing an aggregate governance indicator would disappear since we really only have just a single data source. For his part, K goes so far as to assert that "...this unknown but substantial degree of interdependence among many of the sources also obviates any claims regarding the "precision" of these indicators." (p. 23).[11]

This example is of course contrived because it makes the implausible assumption that the two data sources make perfectly correlated measurement errors

[11] Of course this raises a logical puzzle -- if the degree of correlation in errors across sources is unknown, how can K know that it is "substantial"? Below we discuss in more detail other work we have done which allows us to identify empirically the degree of error correlation across sources, making it "known" -- at least conditional on identifying assumptions -- and also showing that it is in fact not "substantial".

when they assess governance across countries. A first important point to note is that any deviation from this assumption of perfectly correlated errors means that there are in fact gains in precision to be had from aggregation. Even if the errors made by the two data sources are highly, but not perfectly, correlated, an aggregate indicator averaging the two will be at least somewhat more informative than either individual indicator. In short, the presence of correlated errors among sources does not eliminate the benefit of constructing an aggregate governance indicator, although it does of course reduce it. This concern is also not new. In fact, in our very first methodological paper on the aggregate governance indicators (Kaufmann, Kraay and Zoido-Lobatón 1999a) we devoted an entire section of the paper to this possibility, and showed how the estimated margins of error of our aggregate governance indicators would increase if we assumed that the error terms made by individual data sources were correlated with each other. We also note that even if two data sources make correlated errors, it does not mean that we should discard them entirely from the aggregate indicator -- they jointly still might well contain useful information, just not as much information as they would if they were truly independent.

The more important empirical question is whether this correlation of errors across sources is large or not. Both AO and K offer only anecdotal evidence of cases where some of our specific data sources have access to other of our data sources when formulating their assessments. We note first however that the mere fact that data sources may "look at each other" does not by itself constitute evidence that these data sources will therefore make correlated errors. It is useful to think of the assessment of any data source as providing some "signal" of governance, combined with an error term capturing the idiosyncracies of that particular data source. Suppose that one commercial risk rating agency decides to look at the necessarily noisy estimate of governance produced by another rating agency. Surely the first agency, which is in the business of providing informative estimates of governance to its customers, has every incentive to try to filter out the measurement error from the other data source that it is looking at, before incorporating it into its own estimates. While we do not pretend to know exactly how all of our individual data sources process the information at their disposal, it seems strange to us to suggest -- as implicitly do AO and K -- that they blindly copy each other and so make correlated errors.

AO and K both also observe that different data sources might be influenced by the same media reports about a country, and argue that this justifies their claim that individual data sources make correlated errors. Logically this does not follow, as it depends on whether the media reports are accurate or not. If the media reports are accurate, then all the individual data sources that rely on this common media report will both be more accurate, and also more correlated with each other, as a result -- and this surely would be a desirable outcome. Of course, some media reports are more accurate than others, but AO and K do not enter into this crucial part of the argument.

While AO and K both provide some anecdotal evidence of data sources making correlated errors, only K attempts to provide some empirical evidence, for one measure of corruption.[12] K first documents convincing evidence of a methodological break in one widely-used expert assessment of corruption, the International Country Risk Guide (ICRG), in October of 2001, noting that in this particular month an extraordinarily large number of countries in the sample have their scores change when compared with typical other months. He then goes on to point out that, compared with earlier dates, the ICRG corruption ratings become more correlated with the Transparency International (TI) corruption ratings, with the correlation increasing from 0.71 to 0.92. He concludes that this provides evidence that the ICRG corruption ratings are not independent of the other data sources embodied in the TI ratings, and in particular suggests that the reason for the methodological break was to camouflage a move to greater consistency with the TI ratings. While interesting, we do not find this tidbit of evidence to be compelling, for two main reasons.

[12] We note in passing that two of the four examples offered by AO are either incorrect or exaggerated. They incorrectly state that we use the Cingranelli and Richards Human Rights database and the Political Terror Scale, which both rely on numerical coding of information in the US State Department's Human Rights Report, as separate data sources in the same indicator. We do indeed use data from both these sources, but recognizing their common origin in the State Department reports, we average them together and use them as a single indicator in the aggregate indicators. Unfortunately our documentation of this detail in our data appendices was not completely clear. They also suggest that we use three different data sources from Freedom House in the same aggregate indicators. This is in fact the case only for Voice and Accountability. For Rule of Law and Control of Corruption we do rely on two data sources from Freedom House, Nations in Transit, and Countries at the Crossroads. However, in 2005 these two data sources, covering 28 and 30 countries respectively, have just two countries in common, Russia and Tajikistan. This means that there is practically no opportunity for correlated errors between these two sources to have any effect.

- First, the mere fact that the ICRG correlation with TI increases does not provide any evidence at all that the errors made by ICRG are correlated with the errors made by the other sources embodied in TI. Nor does it even provide evidence that the ICRG scores are based on the TI scores. The increased correlation with TI following the methodological break could logically also reflect the fact that ICRG had improved the quality of its own assessments, and improving the signal to noise ratio in its own assessments made it more correlated with other assessments. A purely "home-grown" improvement in quality by ICRG could therefore also account for the higher correlation with TI.

- Second, and perhaps more important, this pattern of increased correlation with other sources following methodological breaks is not in fact a systematic feature of the ICRG ratings. From the standpoint of analysis, it is fortunate that ICRG has in fact made methodological changes to several, but not all, of its indicators, in two different years. In Kaufmann, Kraay and Mastruzzi (2006) we have systematically looked at the 10 specific ICRG indicators we use in the WGI, and identified two series with methodological breaks in 1997, and five in 2001. If the objective of such breaks really were to generate new ratings that are more correlated those of with other experts, as argued by K, then we should systematically expect to see increases in correlations with other expert assessments when comparing the period before and after the methodological break. In contrast, we should see no change in the correlation of ICRG with other expert assessments for series that did not have methodological breaks in the same year. It turns out that this simply is not true in the data. We do find, consistent with K, that the correlation of the ICRG corruption rating with other expert assessments increases after the break in 2001. But when we compare the change in correlations with other expert assessments in the set of five ICRG indicators with breaks in 2001, with the change in correlations of the remaining five ICRG indicators without breaks, we find virtually no systematic difference. In fact, the typical change in correlation of indicators with breaks is just 0.01, while the correlation of indicators without breaks is almost identical at -0.01. If, as suggested by K, ICRG has used methodological breaks to camouflage a greater correlation with other sources, then it is very puzzling why they should

not do so systematically. We therefore do not find K's isolated evidence for just one of the many ICRG indicators to be at all compelling.

Finally, in our latest paper on the governance indicators (Kaufmann, Kraay and Mastruzzi 2006) we have provided new empirical evidence on the possible correlation of errors across data sources. As we discuss at length in that paper, empirically identifying correlations in errors across sources is difficult. Simply observing that two data sources provide assessments that are highly correlated is not enough, since the high correlation could reflect either (i) the fact that both sources are measuring governance accurately and so are highly correlated, or (ii) the fact that both sources are making correlated measurement errors in their assessments of countries. In order to make progress we need to make assumptions, and in that paper we detail two sets of assumptions that allow us to disentangle potential sources of correlation in the errors. One assumption is related to the plausible argument of K that surveys of firms or individuals are less likely to make errors that are correlated with other data sources than, for example, the assessments of commercial risk rating agencies. If this is the case, however, we would expect that the assessments of commercial risk rating agencies be very highly correlated with each other, but less so with surveys. This turns out not to be the case. For example, the average correlation among our five major commercial risk rating agencies for corruption in 2002-2005 was 0.80. The correlation of each of these with a large cross-country survey of firms was actually slightly higher at 0.81, in contrast with what one would expect if the rating agencies had correlated errors. We do this exercise for components of all six of our aggregate governance indicators, and find at most quite modest evidence of error correlation.

<u>Critique 8:</u> If some data sources make correlated errors, the aggregation procedure used by the WGI gives too much weight to such indicators.

The WGI are constructed using a statistical methodology known as an unobserved components model, which in effect estimates governance for each country as a weighted average of the underlying indicators available for that country. The premise for the weighting of indicators is simple. We think of each underlying data source as providing a noisy or imperfect signal of governance. If the errors made by individual sources are uncorrelated with each other, then data sources that produce

highly correlated ratings must be more informative about governance than sources that are not highly correlated with each other, i.e. the variance of the error terms must be relatively small. Accordingly, these more highly-correlated data sources should receive a greater weight in the aggregate indicator. This neat logic however breaks down if the correlation between data sources is due to their making correlated errors, along the lines discussed in the previous critique. If this were the only reason for the correlations among sources, then more highly correlated sources should receive less weight, not more. This point is made by both AO and K. Specifically, given their concerns that the errors made by commercial risk rating agencies are likely to be highly correlated (a claim we have argued above not to be very compelling), they conclude that the WGI place excessive weight on such data sources.

As we have discussed above, isolating the effect of correlated errors in driving the observed correlation among sources is extremely difficult, and so a general and robust solution to this potential problem is not yet within our reach. We acknowledge the point that our weighting scheme could in principle be giving more weight to particular sources for the "wrong" reasons, although as we have discussed at length and in Kaufmann, Kraay and Mastruzzi (2006), we have not yet found compelling evidence of substantially correlated measurement errors in our sources.

Nevertheless, we do accept the point that it is useful to explore how different our aggregate indicators would look if we used alternative weighting schemes.
The simplest approach is to just construct an unweighted average of our data sources. This in practice substantially reduces the weights applied to expert assessments from commercial risk rating agencies, the main concern of AO and K. We have done this for all six of our indicators for the past seven years, and it turns out that the equally-weighted indicators are extremely highly correlated with the benchmark aggregate indicators. The average (across the 42 indicator-year combinations) correlation between the equally-weighted indicators and our benchmark indicators is 0.99. In only three cases is the correlation less than 0.99, and the minimum correlation is 0.97. This clearly shows that equal weighting of our underlying data sources does not practically affect our estimates of governance in the vast majority of cases.[13] The reason for this very high

[13] The main benefit however of weighting sources by their precision is that it yields somewhat smaller standard errors, allowing for more precise inference about cross-country differences and

correlation between the weighted and unweighted aggregate indicators is that by and large, all of our individual data sources tend to agree quite strongly with each other in their assessments of governance.

In fairness, we do think that K suggests a number of reasonable alternative weighting schemes. Motivated by his suggestions, we here consider an alternative weighting scheme that organizes data sources by type and constructs an aggregate indicator weighting each *type* of indicators equally. We consider four groups: (1) cross-country surveys of firms and individuals, (2) commercial expert assessments, (3) expert assessments produced by NGOs, and (4) expert assessments produced by multilateral organizations. We simply average our indicators within each group, and then construct an aggregate indicator based on the four groups alone. We find that the correlation of the resulting indicators with our benchmark indicators is again very high, averaging 0.95 for five of the six indicators in 2005. The only exception to this pattern is for our Regulatory Quality measure, which has a correlation of only 0.44 with the benchmark measure. However, a closer look at the data reveals that this is entirely due to one data source, the Business Environment and Enterprise Performance Survey (BEEPS) carried out the World Bank in transition economies. This source alone among the very many we use, for this particular indicator alone, gives quite different assessments from other types of data sources. If we drop the BEEPS from the exercise, we find that our revised indicator based on four groups of data has a correlation of 0.96 with the benchmark indicator.[14] In fact, this example highlights the risks of relying exclusively on one or another individual data source, which may provide quite idiosyncratic assessments of governance, and points to the usefulness of considering a wide range of data sources, especially as summarized in an aggregate indicator.

To sum up, this critique and the previous one deal with the tricky issue of whether individual data sources on governance do in fact provide fully independent estimates of governance. We have argued that this is in general a difficult problem to

---

changes over time in governance. On average, the standard error of our equally-weighted indicators is about 10 percent higher than in our benchmark indicators.

[14] We note in passing that while in general we think that firm survey data is extremely useful, it can also at times be anomalous, particularly in highly non-market or undemocratic countries. It is striking for example that on the BEEPS survey of firms, Belarus and Uzbekistan rank 4th and 6th *best* out of 27 countries on questions about corruption, while a commercial risk rating agency (DRI) ranks these countries 23rd and dead last, respectively.

sort out, yet in the cases where we can make progress the data do not provide any clear evidence of strong dependence in the errors made by individual data sources. In principle, the main reason to be concerned about the independence of data sources is that it matters for the appropriate weighting of the individual indicators in the aggregate indicator. It turns out, however, that since all of our data sources on average do tend to agree quite strongly with each other, the particular weights used to construct the aggregate indicators do not make a significant difference in the overall estimates of governance.

Critique 9: The WGI lack "construct validity"

Thomas (2006) (hereafter, T) offers yet a different critique of the WGI. Her claim is that the WGI lack what is referred to in a few realms of social science research as "construct validity", and so their use by policymakers is "arbitrary". To understand this rather sweeping (and as we argue, unsubstantiated) critique, which is likely to be unfamiliar to many outside the few particular disciplines that use this concept (and especially unfamiliar to most economists and political scientists), it is useful to provide some background. The idea of construct validity first emerged in psychology in the 1950s as a means of disciplining the process of defining and measuring personality traits in humans.[15] It has since spread to some other disciplines, although notably not to economics, as a tool for judging the usefulness of abstract concepts and their empirical proxies. There are two main ingredients to construct validity. First, a construct, such as "self-esteem" (to take an example from psychology) needs to have a clear definition. Second, for self-esteem to be a useful construct, alternative measures of self-esteem should satisfy (a) "convergent validity" in the sense of being highly correlated with each other, and (b) "discriminant validity" in the sense of not being very highly correlated with other constructs, for example, self-absorption.

T first criticizes the WGI for failing to provide adequate definitions of the six dimensions of governance. For example, she objects to our "Voice and Accountability" indicator because its definition does not coincide with that of another scholar (in this case, Hirschman (1970)), and similar she objects to our "Rule of Law" definition because it does not coincide with another scholar's definition (in this case Fallon (1997)). We

[15]See Trochim (2006) for a very accessible on-line survey article on construct validity

reject this line of criticism entirely as definitional nitpicking. For better or worse, it is widely acknowledged by scholars and practitioners that the term "governance" lacks an accepted common definition. In the absence of such definitional consensus, we think it is entirely reasonable for us to propose our own definitions of the six governance indicators, of course basing them broadly as we did on existing definitions or understandings of the concepts.[16] And for some of our indicators, we think the definitional ambiguities are minimal in any case. Control of Corruption for example is very clearly based on the accepted definition of the term as the use of public office for private gain. In our view, it is completely unwarranted to dismiss these indicators as "elaborate untested hypotheses" (as T rather dramatically does in her abstract) merely due to what in our view are no more than definitional quibbles. In fact, if we were to take T's critique seriously, the absence of definitional consensus in the profession regarding governance would paralyze any effort to measure governance using any means.[17]

T also criticizes the WGI for failing to provide explicit evidence of "convergent" and "discriminant" validity. In our context, convergent validity requires, for example, our alternative measures of corruption to be highly correlated with each other. In fact, the various individual components of our governance indicators are quite highly correlated with each other within each of the six governance indicators. To take an example noted earlier, the median correlation of each of the 18 individual data sources we have for Control of Corruption in 2005 with the aggregate indicator is 0.85! More generally, this high intercorrelation of individual data sources within each aggregate indicator can easily be inferred directly from the margins of error that we report with the estimates of governance. It is straightforward to see that if the pairwise correlations among our individual data sources were low, the estimated margins of error for country estimates

[16] While we do not think that playing definitional "gotcha" is particularly helpful, we do note that our definition, and individual indicators, of rule of law actually correspond quite closely with the definition we recently discovered, offered by another legal scholar. Dakolias (2006) refers to "Rule of Law" as the presence of "...meaningful and enforceable laws where decisions are transparent, fair, and predictable; enforceable contracts that promote business and commerce; basic security with personal safety; protection of individual and property rights, and an independent judiciary that safeguards both; and access to justice with concrete ways to invoke rights and protect them."

[17] We note also that there is "free entry" in the market for governance indicators, and if T, or any other scholar, would like to provide an alternative definition of some dimension of governance, and provide an empirical measure of it, then they can certainly do so. We would at a minimum expect a serious critique of our indicators to provide (i) provide a clearly superior alternative definition, (ii) operationalize it in the form of a new indicator of governance, and (iii) show that it differs significantly from what we have produced. Regrettably, T does none of these three things.

would approach infinity. While we have long noted that the margins of error associated with our estimates of governance are non-trivial, they are also not enormous, and this by itself demonstrates "convergent" validity.[18]

"Discriminant" validity requires our empirical proxies for, say, corruption, not to be correlated with any other dimensions of governance. We do not think that, in the case of measuring governance across countries, the concept of discriminant validity is at all useful in assessing empirical measures. The reason is very simply that different and logically distinct dimensions of governance are in reality likely to be correlated across countries, and so the individual proxies that we use to measure them will also be correlated. For example, one can readily think of causal mechanisms through which the institutions of democratic accountability (that we try to capture through our Voice and Accountability indicator) help to reduce the incidence of corruption.[19] Since the two "constructs" of Voice and Accountability, and Control of Corruption, are likely to be correlated with each other in reality, it is futile to expect that individual indicators of the one will be uncorrelated with individual indicators of the other. This suggests to us that discriminant validity simply is not a useful criterion in this context. In fact, applying strictly the criterion of discriminant validity would necessarily rule out *any* measures of these two concepts.[20]

---

[18] While we argue that the machinery of "construct validity" is not especially useful in our application, it is interesting to note that the statistical procedure that we use to construct the aggregate indicators in a sense rewards indicators that display greater "convergent validity". As noted in the previous critique, individual indicators that tend to be more highly intercorrelated are assigned greater weight in the aggregate indicator.

[19] This is an example of what T refers to as "functional relationships" between various dimensions of governance, that T asserts need to be fully modeled before governance can be measured – an assertion we clearly disagree with.

[20] The lack of usefulness of the notion of discriminant validity is even clearer at a higher level of aggregation. T suggests that, because the six aggregate indicators in the WGI are highly correlated with each other, that they are invalid because they do not display discriminant validity. We fail to see how this constitutes a meaningful critique of our indicators. For example, democratic accountability and corruption are logically distinct concepts, but the notion of discriminant validity would suggest that they cannot be valid concepts *simply because they happen to be, for good reason, correlated across countries*. Construct validity rapidly descends into a spiral of circular reasoning from here. Perhaps a way to justify "discriminant validity" in the case of correlated concepts is to ask whether the correlation of an individual measure of democracy with an individual measure of corruption is "low enough" given the correlation that exists in reality between democracy and absence of corruption. But we cannot know what is "low enough" without knowing the true correlation between democracy and corruption. And this we can't know without empirical proxies!

Of course, at a broader level, we do accept that reasonable people can differ on what constitutes an appropriate definition of governance and its key dimensions. Reasonable people can also disagree with our mapping of individual measures of governance to the corresponding aggregate indicators, and we realize that different people might prefer both a different set of aggregate indicators and/or a different mapping of individual indicators to the aggregate indicators.[21] However, we fail to see how such legitimate differences of opinion would invalidate the aggregate indicators that we have proposed (and that are now widely used) -- rather, the critique might just more modestly suggest that for certain purposes, other measures of governance might be appropriate, and we certainly would not dispute this.

Finally, we categorically reject T's concluding claim that the use of our indicators by policymakers is "arbitrary" because of the supposed failure to establish "construct validity". For policymakers grappling with the difficult problem of trying to measure corruption or other dimensions of governance across countries, failure to rely on available data would be arbitrary in the extreme. Endlessly waiting for the articulation of a complete, coherent and consistent theory of governance before proceeding to measurement and action (of course with due regard for data limitations), while perhaps intellectually satisfying to a few, would be impractical to the point of irresponsibility.

Critique 10: The WGI project is insufficiently transparent

This critique is raised most strongly by T, but also to some extent by AO, and is mostly concerned with access to the data from the underlying sources on which the WGI are based. T goes so far as to assert that "...the data upon which they [the WGI] are based are not available to the research community to allow evaluation, critique or

[21] We nevertheless strongly disagree with the discussion around Box 1 in T, where she presents an entirely contrived example to illustrate her claim that the ranking of countries is significantly affected by how we assign individual indicators to the aggregate categories. A quick glance at T's example shows that the reversal of rankings under alternative clustering schemes is purely an artifact of the fact that (a) the hypothetical individual indicators she presents use different units, and (b) the individual indicators are essentially uncorrelated with each other. The example is therefore entirely uninformative about the assignment of indicators in our work, where the indicators are rescaled to common units and are highly correlated with eachother. In order for T's critique on this point to have any content whatsoever, it needs to move beyond trivial hypothetical examples to taking the actual data that we use (and that we publicly provide on our website), to see whether reasonable re-groupings of our individual indicators lead to substantively different country rankings.

refinement", and as a consequence we have left "...replication and peer review as impossibilities."

With one important caveat we think this line of criticism is entirely unfounded. First, the caveat. Several years ago we decided to include the Country Policy and Institutional Assessment (CPIA) ratings from the World Bank, African Development Bank, and Asian Development Bank among the sources for the WGI. We included these data sources based on our judgement that they provided valuable information on the dimensions of governance we wanted to measure. Unfortunately, however, only as of 2005 have these institutions fully publicly disclosed the CPIA data, and then only for the set of low-income countries eligible for concessional lending from the International Development Association (IDA). For a few years earlier the CPIA data was disclosed, but at a more aggregated level and only by quintiles, for the same set of IDA-eligible countries. And prior to this the CPIAs have been confidential, which we have had to respect.

We have however fully disclosed all of the remaining data used to construct the WGI. Much of this data is already in the public domain, and we have simply reproduced the data as we have used it in the WGI simply for the convenience of users. The data from our proprietary data sources is commercially available to anyone, and moreover these data sources have kindly agreed to allow us to post their proprietary data, again exactly in the way that we use it in the WGI. This means that users of the WGI can visit our website (www.govindicators.org) and immediately retrieve the data from our individual sources exactly as it enters into the WGI.

T's critique of the availability of our data for replication and peer review can thus only be based on the fact that, in some cases where we average individual questions from a single data source before including it in our aggregate indicators, we have thus far not posted on our website the data at this lowest level of aggregation.[22] We think this concern is marginal. This critique applies only to six of our commercial data sources that

[22] Our primary reason for posting the data at this level of aggregation is because we thought it would be most useful for helping users understand how levels and changes in our aggregate indicators depend on changes in the underlying data sources. If there is sufficient demand (which so far has been negligible) we may subsequently explore the possibility of posting the fully disaggregated data.

are not elsewhere available in the public domain in some form already. Of these six, in many cases we do not average questions anyhow, and so there is no further level of disaggregation to disclose.[23] In fact, for the Control of Corruption indicator, which is by far the most widely-used of six WGI measures, there is not a single case where the data is not freely publicly available at the lowest level of disaggregation, either through our website or elsewhere.

We thus find T's concerns about access to underlying data, and the consequences for scrutiny and peer review, to be vastly exaggerated. With the important partial exception of the CPIA data noted above, all of the data used to construct the WGI is in fact publicly available on our website and can be used for replication and robustness testing. In those cases where users might for some reason want to further disaggregate the data from individual data sources, all such data are available, in most cases freely although sometimes on a commercial basis. We note also that our level of data disclosure exceeds accepted standards in the economics profession (again of course with the unavoidable exception of the confidential CPIA data).[24] We certainly agree with T about the importance of public scrutiny of all types of governance data. Our personal views are that the usefulness and credibility of the CPIAs would be enhanced if they were fully public for all countries and all years. We are also concerned that a number of new initiatives of international organizations to measure public sector accountability, such as the Public Expenditure and Financial Accountability (PEFA) initiative and the OECD-Development Assistance Committee's assessments of public sector procurement practices are proceeding apace with insufficient prior guarantees that the resulting data will be made public.[25]

---

[23] This applies to roughly 40 percent of the country-year observations from our main commercially-available data sources.

[24] The two most prestigious general-interest journals in economics, the American Economic Review and the Journal of Political Economy, share the same explicit data disclosure policy (see http://www.aeaweb.org/aer/data_availability_policy.html and http://www.journals.uchicago.edu/JPE/datapolicy.html). The general principle is to post data in order to allow replication of reported results, which is what we have done. Beyond this, in the case of commercially-available datasets, the policy requires only that authors provide instructions on how to obtain such data, as long as third parties are in principle able to obtain the same dataset. We have gone well beyond this standard by disclosing the commercial and non-commercial data required to replicate our results. We therefore reject T's claims that we have not made our data sufficiently available for replication and peer review, at least by the accepted standards of our profession.

[25] For example, the PEFA indicators have been constructed for roughly 30 countries as of end-2006, but in only nine cases are the data publicly available.

<u>Critique 11</u>: Flaws in the evidence on the two-way relationship between governance and growth in Kaufmann and Kraay (2002)

We conclude with a brief discussion of the penultimate section of AO, where they critique an empirical exercise we carried out in Kaufmann and Kraay (2002) using one of the WGI, the Rule of Law measure.[26] In that paper we were interested in providing a first set of estimates of the reverse causation from income levels to institutional quality. In order to identify the causal effect running from institutions to per capita incomes, we borrowed the framework of the famous paper by Acemoglu, Johnson and Robinson (2001), who developed a clever instrument for institutions based on historical settler mortality rates. A puzzling feature of that paper, as well as our replication of it with somewhat different data, is that the estimated causal effect of institutions on per capita incomes was much larger than the slope coefficient in a simple OLS regression. We then showed how with the help of a few minimal assumptions about the degree of measurement error in institutional quality and the role of omitted variables we could pin down the magnitude of the reverse effect, from per capita incomes to governance, in a way that was consistent with the puzzling relationship between the OLS and IV coefficient estimates in Acemoglu et. al.. Our somewhat surprising conclusion was that, for reasonable assumptions, this reverse effect was negative, implying that exogenous shocks to income could worsen governance.

AO present an extension of this exercise, where their main innovation is to propose a new strategy that allows them to directly estimate the causal effect of per capita incomes on institutional quality. Their strategy consists of proposing an instrument for per capita incomes. In particular, AO, and we, consider a two equation system in governance and per capita incomes. As usual, identification requires the presence of a variable that affects governance but has no direct effects on incomes (the famous settler mortality instrument), and another variable that affects per capita incomes but has no direct effect on governance other than through its effect on income. Finding such a second instrument is very hard -- sufficiently so that we undertook to write our

[26] We note in passing that KS also provide a critique of the governance and growth literature that we find to be entirely without merit. Our views on their critique can be found in the response we have prepared for the Journal of Politics.

'Growth without Governance' paper proposing a strategy for avoiding such a difficult task.[27]

Unfortunately, we do not think that the instrument proposed by AO is especially compelling. They use infant mortality rates as an instrument for per capita incomes, arguing that infant mortality is unlikely to have any direct links to institutional quality. Yet this seems hard for us to believe on *a priori* grounds. Infant mortality rates depend in considerable measure on public health interventions, and it seems plausible to us that the quantity and quality of these in turn depend at least in part on governance. Or mechanically, infant mortality rates are quite likely to be significantly correlated with historical settler mortality rates, and these in turn are correlated with our measure of institutional quality. Thus it seems implausible to us that their proposed instrument satisfies the exclusion restriction.

We certainly do not want to claim that our paper is the "last word" on the magnitude of the feedback from incomes to governance, and we strongly agree that more work is needed in this area. Unfortunately we do not think the particular strategy adopted by AO is a very fruitful one.

Conclusions

It is useful to summarize the various critiques by noting that several of them, most notably Critiques 1, 2, and 4 that deal with comparisons of governance over time and across countries, we think are primarily based on a misunderstanding of our aggregate indicators and their interpretation. Several other critiques, notably Critiques 3, 7 and 8 on error correction and correlated errors on the part of commercial risk rating agencies, we acknowledge are conceivable on *a priori* grounds. However, we have argued that these critiques either are entirely lacking in empirical support, or even if they are empirically supported to some extent, the effects are so small as to be practically irrelevant. The same is true for Critique 5 dealing with halo effects, which again are of course possible, but we have yet to see convincing evidence of their empirical importance, and the available empirical evidence so far is simply not robust. Finally, we

[27] Rigobon and Rodrik (2004) provide an alternative clever identification strategy of this two-way relationship that also does not rely on searching for a good instrument for per capita incomes.

have argued that Critique 9 regarding "construct validity" involves the application of a framework that simply is unsuited to the task at hand.

In conclusion, we reiterate that we think the process of debating and questioning the many conceptual and methodological required to construct indicators of governance is a very useful one. We appreciate the effort that has recently gone into producing written critiques of the WGI project as it is essential to sharpening our understanding of the issues involved with measuring governance. We also do not want to suggest that the WGI project has provided the best possible measures of governance. There are many areas where progress in measuring governance across and within countries is both valuable and possible, and we hope that work in these areas continues. But at the same time we emphasize that many of the concerns raised by our critics are not specific to the WGI, and are likely to arise in the context of many other existing and proposed efforts to measure governance. The relevance of these critiques should therefore also be carefully considered by the producers of other measures of governance as well.

References

Acemoglu, Daron, Simon Johnson, and James Robinson. "The Colonial Origins of Comparative Development. *American Economic Review*. 91(5):1369-1401.

Arndt, Christiane and Charles Oman (2006). "Uses and Abuses of Governance Indicators". OECD Development Center Study.

Dakolias, Maria (2006). "Are We There Yet?: Measuring Success of Constitutional Reform". Vanderbilt Journal of Transnational Law. 39(4):1117-1231.

Evans, Peter and James Rauch (1999). "Bureaucracy and Growth: A Cross-National Analysis of the Effects of 'Weberian' State Structures on Economic Growth". *American Sociological Review* 64 (October 1999): 748-765.

Glaeser, Edward, Rafael LaPorta, Florencio Lopez-de-Silanes, and Andrei Shleifer (2004). "Do Institutions Cause Growth?". Journal of Economic Growth.

Kaufmann, Daniel, Aart Kraay and Pablo Zoido-Lobatón (1999a). "Aggregating Governance Indicators." World Bank Policy Research Working Paper No. 2195, Washington, D.C.

Kaufmann, Daniel, Aart Kraay and Pablo Zoido-Lobatón (1999b). "Governance Matters." World Bank Policy Research Working Paper No. 2196, Washington, D.C.

Kaufmann, Daniel, Aart Kraay and Pablo Zoido-Lobatón (2002). "Governance Matters II – Updated Indicators for 2000/01." World Bank Policy Research Working Paper No. 2772, Washington, D.C.

Kaufmann, Daniel and Aart Kraay (2002). "Growth Without Governance". Economia.

Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2004). "Governance Matters III: Governance Indicators for 1996, 1998, 2000, and 2002". World Bank Economic Review. 18:253-287.

Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2005). "Governance Matters IV: Governance Indicators for 1996-2004. World Bank Policy Research Working Paper No. 3630. Washington, D.C.

Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2006). "Measuring Governance Using Perceptions Data", in Susan Rose-Ackerman, ed. Handbook of Economic Corruption. Edward Elgar.

Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2006b). "Governance Matters V: Governance Indicators for 1996-2005. World Bank Policy Research Department Working Paper No. 4012.

Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2006c). "Measuring Corruption: Myths and Realities". Development Outreach. September 2006, pp 37-41.

Knack, Steven (2006). "Measuring Corruption in Eastern Europe and Central Asia: A Critique of the Cross-Country Indicators". World Bank Policy Research Department Working Paper 3968.

Razafindrakoto, Mireille and Francois Roubaud (2006). "Are International Databases on Corruption Reliable? A Comparison of Expert Opinion Surveys and Household Surveys in Sub-Saharan Africa". Manuscript, IRD/DIAL, http://www.dial.prd.fr/dial_publications/PDF/Doc_travail/2006-17.pdf.

Thomas, Melissa (2006). "What Do The Worldwide Governance Indicators Measure?" Manuscript, Johns Hopkins University.

Trochim, William (2006). "Construct Validity", available at: http://www.socialresearchmethods.net/kb/constval.htm